The Stability Problem of Value-added Models in Teacher Effectiveness Estimations: A Systematic Review Study

Ismail Aslantas

School of Education, Durham University, Durham, United Kingdom

ismail.aslantas@durham.ac.uk

The Stability Problem of Value-added Models in Teacher Effectiveness Estimations: A Systematic Review Study

This article provides evidence by undertaking a systematic review on the stability problem of using value-added models in teacher effectiveness estimates from the perspective of the impact of the number of previous test scores employed aimed at answering a unique review question (This is a long sentence that lacks clarity – it could be broken down for a clearer meaning): How stable is teacher effectiveness estimates measured by VAMs? By using the terms: teacher performance, student performance, value-added model, stability and their other related synonyms, a comprehensive search was conducted in 17 databases along with employing hand search in Google Scholar and contacting authorised persons by email. In total 1439 records were found as a result of the searches. After completing the screening process, 50 studies remained for data extraction.

Out of 50 a total of studies in the review list, 13 focused on the stability of VAM estimates regarding using the number of prior test scores. In summary, there is a common view that the use of prior year data in on value-added estimates for teacher effectiveness has a positive impact, however, with regard to the impact of multiple previous year data, different voices arose from the researchers.

Keywords: teacher effectiveness, value-added model, stability, systematic review

Introduction

Although the use of value-added models (VAMs) in the field of business and economics at first, as researchers in the education sector began to be interested in these models to measure teacher performance, it appears that there were research (studies) on VAMs in this field in the late 1900s. (Sanders & Horn, 1994, 1998; Sanders, Saxton, & Horn, 1997; Webster & Mendro, 1997; Kupermintz, 2003). With the effect of No Child Left Behind Act of 2001 (NCLB) based on the belief that the most important school-related factor affecting student achievement is the quality of the teacher (Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005), the studies done in this field gained momentum (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louris, & Hamilton, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010). In order to measure the effectiveness of teachers, several types of VAMs were developed and applied by states and school districts in the US such as the Tennessee value-added assessment system (TVAAS) and the Dallas value-added accountability system (DVAAS). The common assumption of these models that the growth/decline of the students' achievement in the standardised tests is attributed to their teacher performance. In all VAMs concepts, it is assumed that the quality of students' performance in the school reflects the quality of the teaching received (Darling-Hammond, 2015). A teacher's value-added scores can be calculated by subtracting his/her students' predicted scores from their actual scores in the standardised test(s) (Sanders, Saxton, & Horn, 1997). The predicted score is subjected to the same students' one or more previous year test scores and their characteristic features in most VAMs (Ouma, 2014). These prediction methods are likely to be beneficial to identify the most and least effective teachers in a school, district and/or state (Colorado, 2007).

As a result of the increasing number of research studies in parallel with the interest of researchers in this field, it was possible to discover the strengths and weaknesses of VAMs. The teacher performance results estimated by VAMs are highly affected by students' characteristics and other conceptual predictors which teacher cannot control (Wei, Hembry, Murphy, & McBride, 2012). These models should not be used for the high-stake decision about the teachers unless the pros and cons of VAMs are fully revealed. Therefore, their use should be limited to improve the educational institutions, provide teachers with the opportunity to address their own shortcomings and provide justification for students' academic progress. More accurate information is needed about value-added models to expand the intended use of them. One of the robust methods to collect accurate information from the literature is conducting a systematic review. Primarily the research has the aim to determine whether the teacher performance results estimated by applying value-added models under different conditions are stable or not. Specifically, observable teacher characteristics, school characteristics, the students' test scores obtained over time and the preferred data analysis methods are the different conditions that are referred to in this current research. As one of the purposes of this study is to provide guidance and appraise to policymakers and practitioners on the use of value-added models in the teacher performance appraisal for high-stake purposes such as decisions on dismissal and monetary reward, gathering evidence-based findings from a wide body of research systematically are needed instead of from narrative literature. Therefore, this systematic review is utilised in order to

synthesise the results of previous relevant studies that analysed the impact of conceptual predictors and data analysis methods used on teacher performance evaluation. In line with the purpose of this systematic review declared, a unique review question formulated to assist in exhaustively examining the available evidence is;

How stable is teacher effectiveness estimates measured by VAMs?

where teacher effectiveness operationally defined by VAM as the estimation of the differences between expected and observed student test scores (Kersting, Chen and Stigler, 2013). Moreover, in this systematic review study, the operational definitions of the term stability refer to the stableness of the estimates due to (a) *the number of test scores used*, (b) *the predictors used in the estimations*, and (c) *the analysis methods applied*. The existing literature on the stability of VAMs estimates will be retrieved from these three perspectives.

Methods Databases and Searching Strategy

In order to conduct a comprehensive search in the impact of conceptual predictors and data analysis methods preferred on teacher performance evaluation estimates, both published and unpublished studies that met the inclusion criteria as explained below, are obtained until the 1st of May 2019. In order to identify the studies met the inclusion criteria, a total of 17 electronic databases and six their providers were employed (shown in Table 1). For reaching other relevant sources, research centres, foundations, and researchers who have worked on teacher performance evaluation based on VAMs were contacted personally.

Table 1. *Databases and their providers*

	Provider	Database
1	ProQuest	ProQuest Dissertations & Theses Global: Social Sciences
		Education Database
		ERIC
		International Bibliography of the Social Sciences (IBSS)
		Social Science Database
		Applied Social Sciences Index & Abstracts (ASSIA)
2	EBSCOhost	OpenDissertations
		British Education Index
		Business Source Premier
		Education Abstracts
		Educational Administration abstracts
		PsycINFO
3	Web of Science	Web of Science Core Collection
		Current Contents Connect
4	Elsevier	SCOPUS
5	SAGE Research Methods Core	SAGE Journals
6	Taylor & Francis Online	Educational Research Abstracts Online

To formulate the searching strings, first of all, the keywords which are “teacher performance”,

"student performance", "value-added model" and "stability" were identified in parallel with the review question. Then the related terms of the keywords were determined by identifying which alternative terms were used to substitute the search terms in the related sample studies found (shown in Table 2).

Table 2. *Search keywords*

Search Terms	**Related Terms**	
Teacher performance	Teacher effect*	Teacher proficiency-rank
	Teacher evaluation	Teacher judgment
	Teacher performance evaluation	Educational effectiveness
	Teacher appraisal	Educator performance appraisal
	Teacher performance appraisal	Educator performance
	Teacher quality	Educator evaluation
	Teacher assessment	Educator quality
	Teacher performance assessment	Teaching effect*
	Teacher accountability	Measuring teach*
	Teacher proficiency	Evaluating teach*
Student performance	Academic achievement	Achievement
	Academic gains	Achievement measures
	Student test score	Outcomes
	Student test score	Outcome measures
	Student test-score	Student test performance
Value Added Model	Value added modelling	VAM*
	Value-added model*	Value added estimate*
	Teacher value-added	Value-added estimat*
Stability	Concord*	Imprecision
	Robust	Variat*
	Sensitivity	Fluctua*
	Instability	Persistence
	Precision	Shrink*

Inclusion and Exclusion Criteria

The research included in this systematic review met all criteria located in Table 3.

Table 3. *Inclusion criteria*

Inclusion Criteria	
Criteria	**Description**
The population of this study is teachers evaluated by their students' outcomes	Only studies focused on teacher performance evaluation based on student test scores will be included in this systematic review. Among the studies that are interested in the performance evaluation on more than one subject such as teacher and school performance estimated together in a single study, only the studies whose one of the areas of interest is teachers will be given a place in this systematic review.
The issue of the study is the stability of the estimates	The operational definitions of the term of stability in this systematic review refer to the stability of the estimates; (a) based on students' test scores Studies that use this measure for stability must have at least two previous years test scores for the same or different cohorts of students over time) (b) based on predictors used (c) based on the analysis methods applied Studies are included if they use any one of the above measures of stability.
Only empirical studies are reviewed for this study	Empirical studies refer to primary research as opposed to secondary research such as reviews, government reports. Studies analyzing secondary data such as panel, administrative data are considered as primary research.
The study setting of the research interest is K-12.	(K) refers to kindergarten grade (age 5-6, equivalent to Year 1 in the UK) and (12) refers to the 12^{th} grade (age 17-18, equivalent Y13 or 6^{th} form in the UK. All studies conducted from kindergarten to the 12^{th}grade setting are included in this systematic review.
Published in English	Studies reported in English

The studies were screened with regard to the below criteria as the exclusion criteria of this research;

Not reported or published in English

Not primary research

Not about education

Not within K-12 (e.g. higher education, reception year or nursery)

Not about the evaluation of teacher effectiveness

About the use of value-added measures of teachers to predict teacher attrition

The outcome is not student test scores or gains (e.g. children's behavior or attendance)

Using measures of teacher effectiveness to predict outcomes

Just about school effectiveness or school improvement (the studies focused on both school (principal) and teacher effectiveness, they have potential to be included in the review)

About teacher effectiveness in non-mainstream school

Just about pupils with special educational needs (SEN)

About theories and policies, opinion pieces, discussion pieces

Instructional manual or promotional literature about how to measure teacher effectiveness

Literature about the characteristics of effective teachers

Findings

The research findings from this systematic review are presented mainly in the two sections: description results and thematic analysis results. In the first section, overall explanatory results for the searching process were exhibited. On the basis of the conceptual definition of the stability of VAMs estimates, three themes have been determined for this study; a) the number of previous test scores employed, b) the predictors used, and c) the data analysis methods applied. As this article is a part of the uncompleted doctorate thesis, the key findings of the included studies were only placed under the first theme; the number of previous test scores employed.

Descriptive Results

For a more efficient and well-organized search process, the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) flow diagram is mainly guided. PRISMA is a diagram that published in 2009 by the PRISMA group (Moher, Liberati, Tetzlaff, & Altman, 2009) to help researchers to map out the number of studies identified, included and excluded based on the criteria established. The number of studies included and excluded in the review list was illustrated in the flowchart in Figure 1.

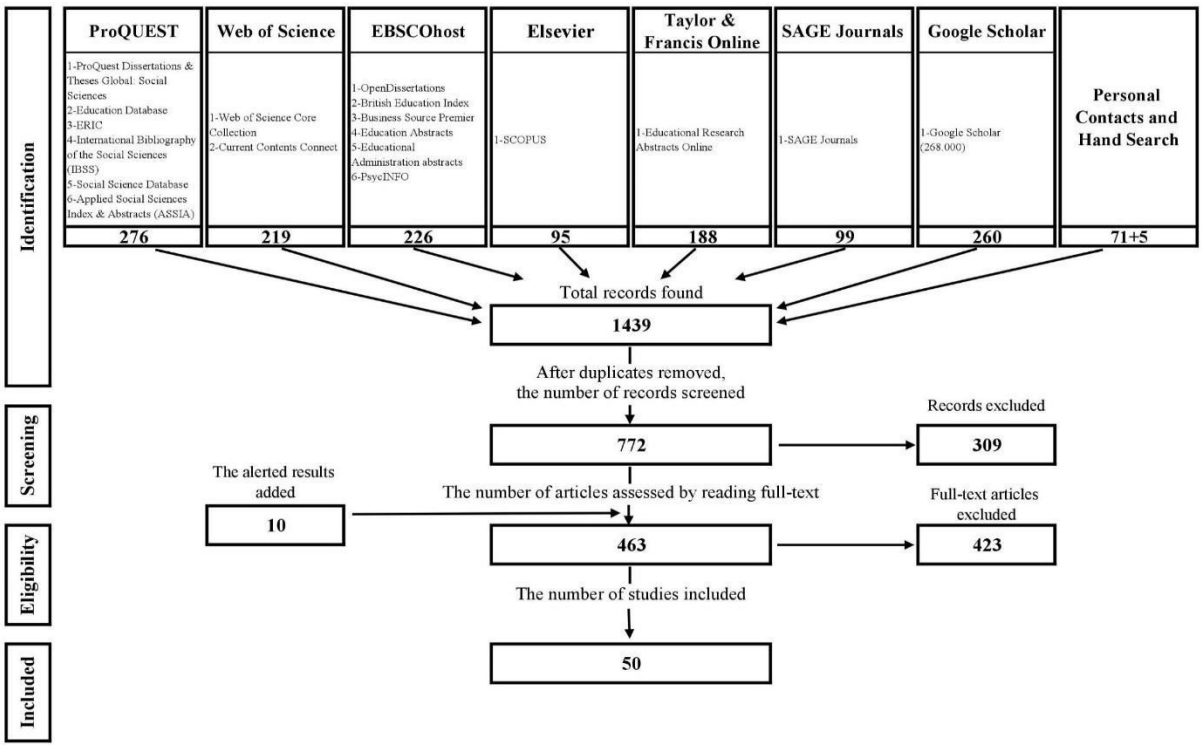

Figure 1. *The Modified PRISMA Flow Diagram*

As a result of typing a combination of the searching strings into the providers' search engines, a total number of 1439 articles were found initially. In the process of merging, 667 cases were deleted from the review list as being duplicated cases of unique studies. Through the phase I screening checklist form, all remaining 772 studies' titles and abstracts were screened, and 309 of them were excluded from this systematic review study in this screening stage. The further 423 cases out of the remaining 473 studies, which were added 10 studies came from the alerted results on the providers, were removed by employing the phase II full-text screening checklist form. To ensure that the screening processes were undertaken away from the prejudice of the researcher and to minimise the lack of potentially relevant articles among the discarded papers, the randomly selected 70 papers were also screened by a second independent reviewer. Although over 95 percent exact agreement reached between the reviewers, the inter-rater reliability agreement was .86.

Finally, the remaining studies were also subjected to the quality appraisal in order to include the findings from only the credible studies in this systematic review. Unfortunately, the trustworthiness of all research done is not the same, so the credibility of their findings should not be the same either. Therefore, to avoid an invalid and misleading conclusion, as of being a practical way for evaluation of the quality of individual studies instead of complex technical checklists in the literature, the modified "sieve" quality appraisal framework designed by Gorard (2014) was employed in this systematic review. As the qualifications of the studies were satisfactory for inclusion in this study, none of them was excluded from the review list. Therefore, by the end of this searching and elimination strategies, 50 research that focused on the stability of teacher effectiveness measurement estimates by VAMs were included in this current study.

Thematic Analysis Results

The final 50 studies retrieved in this systematic review by three main themes; the number of previous year test scores employed, the prediction used, and the data analysis methods applied in estimating teacher effectiveness. Although it was planned to present the results under three themes, as the analysis process in other themes is not yet complete, only the key findings related to the fist theme, the number of previous test scores employed, were placed.

The Number of Previous Test Scores Employed

In this section, the stability of teacher performance evaluation estimations with regard to the number of students' previous test scores used in VAMs was investigated. Although out of 50 studies, a total of 15 research studies are located in this theme.

The first study (Rothstein, 2009) claimed that the limited positive effect of using additional scores was longitudinal research involved 49.456 students from grade 3 to 5 linked to 2844 teachers. The researcher discussed a bias problem in value-added estimates related to student-teacher allocation issues. Although the study covers other aspects of VAM estimates, in this systematic review, only findings related to the impact of a number of test scores from prior years were included. To investigate the impact of using additional year test scores on teacher effectiveness in grade 5, first, the authors added the nearest prior year score (test score in 4^{th} grade) and estimated the increase in R^2, which was .55. The 4^{th}-grade test score contributed to 55 percent to the explanation of variance on test scores in grade 5. Moreover, the authors also included two prior year test scores (pre- and post-tests in grade 3) in the previous model; however, their contribution was very limited, R^2 only increased by .039 (3.9%).

The limited effect reported by the other study (Cunningham, 20014) was a thesis defended for the Doctor of Philosophy degree. In order to provide sufficient information to policymakers and practitioners in making the high-stakes decision regarding with teacher accountability system, this study evaluated the impact of growth model preference, of how many years of data used, and of student-level variable employed in the teacher performance estimates. With using up to three successive years of student data, teacher effects were estimated from five value-added models. Teacher rank orderings obtained in the five value-added models by using either one or three years of test scores were highly correlated with each other. The correlation exceeded .90 when using single-year data, and .80 when conducting multiple years of data. The use of a single year of test scores instead of three years resulted in a slight increase in correlation between the models and a slight decrease in teacher movement between quarters.

The next study, Shafer et al. (2012), compared six growth models used for teacher effectiveness estimates in the literature; quantile regression (QReg), ordinary least square (OLS), growth score difference (year two minus year one), and three different transition models (value-tables). Although the study compares the six growth models, it also covers the impact of using additional previous year data in estimates of the growth models. With related to this theme's concept, the findings on the correlations between scores estimated in reading and maths across four student cohorts allocated were presented. Mainly, the study claimed that the inclusion of data from more previous years (at least two years) in the QReg and OLS

models had a limited positive effect. For instance, the correlation between maths and reading scores for cohort 1 students was found as .19 in QReg1 (used one prior year only) and .18 in QReg2 (used two prior years); similarly the greatest change in correlation coefficients with regard to use additional prior year data in OLS models was found as .01.

Oppositely, eight research in the review list claimed the advantages of using additional prior years data in the teacher effectiveness estimates. One of them done by Lash et al. (2016) in order to investigate how stable the teacher growth percentile scores over the years, so that the authors compared the reliability of coefficients of the estimations. They claimed that stability of performance scores increased from .5 to .67 when the results are obtained by averaging over two years, and to .75 by averaging three years in maths, similarly in reading the increase obtained from .41 to .58 by averaging two years, and .68 by averaging three years.

The next study claimed a substantial improvement obtained in the reliability of VAM estimates by employing additional years of observations is belonging to Goldhaber and Hansen (2013). In order to examine the long-term stability of teacher effectiveness, the authors used up to ten years of longitudinal data in maths and reading across 3rd to 5th grade. After running a series of value-added estimates, the authors reached that there is a substantial improvement in the reliability of the estimates by using multiple prior year test scores. The reliability coefficient increased from 0.29 with a one-year VAM to 0.52 with a six-year VAM.

Another longitudinal study (Hu, 2015) involved 1210 maths, and 1239 reading teachers were conducted to explore the impact of adding student prior achievement into the estimates of teacher effectiveness by using the longitudinal students' data (one to three previous year test scores depend on grade and year) into hierarchical linear models. As a result of this study, students' up to three prior years test scores explained more than half of variance in their current scores was found. Besides, the researcher claimed that not surprisingly, the nearest previous year test score had an important role in this explanation. The average of 57% and 59% of the variance in students' current achievements in maths and reading, respectively, were explained by the nearest prior test scores. Similarly, the additional previous year test scores also contributed to the explanation of the variance in students' current achievement, but not as large as the nearest prior year's. For instance, 67% of total variance in students' mathematics achievement in Grade 7 in 2009-10 were explained by their achievement in grade 6 and 5 (the impact of the achievement score in grade 5 was 12%), and 69% of variance in grade7 in 201011 was explained by the achievement scores in grade 6, 5 and 4 (the impact of achievement score in grade 4 was 2%).

Discussion and Initial Conclusion

This systematic review study is utilised as a part of the researcher's own doctorate thesis in order to synthesise the results of previous relevant studies that analysed the impact of conceptual predictors and data analysis methods used on teacher performance evaluation In line with the purpose of this systematic review, a unique question was formulated in order to retrieve the available evidence. Namely, how stable is teacher effectiveness measured by VAMs? In this systematic review study, the operational definitions of the term stability refer

to the stableness of the estimates due to (a) the number of test scores used, (b) the predictors used in the estimations, and (c) the analysis methods applied. The existing literature on the stability of VAMs estimates will be retrieved from these three perspectives.

Since the researcher's doctoral process is ongoing and the analysis chapter is not completely finished, yet, only the key findings related to the first theme, *the number of previous test scores employed*, were placed in this conference proceedings paper. The stability of teacher performance evaluation estimations with regard to the number of students' previous test scores used in VAMs was investigated. Out of 50 studies, a total of 15 research studies are retrieved in this theme. In general, although there is a consortium on the impact of prior year data on value-added estimates for teacher effectiveness, unfortunately, this consortium is disintegrated about the impact of using additional previous year(s) data on the estimates. Keep in mind that the evidence in this theme is not very robust because of preferring not strong design for their research questions and involving a considerable amount of missing data. Eight studies in this theme, seven of them were rated 2 * from middle bound, and one with 1 * from the lower bound, claimed to be advantageous with adding additional year test scores to the estimates. Although the other seven studies, one of these was rated 2 * from upper bound, and the rest were rated with 2 * from middle bound reported that using additional prior year test scores have a positive impact, but the research also found that the impact is limited, or even little if any. Therefore, the findings are mixed with almost an equal number of medium quality studies suggesting that there are advantages in including additional year test scores as well as those advocating having little benefit. However, the stronger study (rated 2a) suggests that there is little benefit of using additional test scores from previous years. More robust studies may be needed to confirm the results, but at the moment, there is no evidence that using additional prior test scores is useful.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teacher and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Ballou, D., Sanders, W. L., & Wright, P. (2004). "Controlling for Students' Background in value-added assessment for teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Colorado, G. (2007). *An empirical sensitivity analysis of value-added teachers' effect estimates to hierarchical linear model parameterizations.* University of Northern Colorado.

Cunningham, P. L. (2014). *The effects of value-added modeling decisions on estimates of teacher effectiveness.* The University of Iowa.

Darling-Hammond, L. (2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher, 44*(2), 132-137.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy, 1*(1), 19-27.

Goldhaber, D., & Hansen , M. (2010). *Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions.* The Urban Institute. 2100 M Street NW, Washington, DC 20037: National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589-612.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics, 110*, 47-59.

Hu, J. (2015). *Teacher evaluation based on an aspect of classroom practice and on student achievement: A relational analysis between student learning objectives and value added modeling.* Boston College.

Johnson, M. T., Lipscomb, S., & Gill, B. (2015). Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables. *Journal of Research on Educational Effectiveness, 8*(1), 60-83.

Kersting, N., Chen, M.-k., & Stigler, J. (2013). Value-added teacher estimates as part of teacher evaluations: exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7).

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein Critique. *Education Finance and Policy, 6*(1), 18-42.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model* (REL 2016–104) . Institute of Education Sciences, Department of Education. Washington, DC: USA: National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (n.d.). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572606.

McCaffrey, D., Lockwood, J., Koretz, D., Louris, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 647-101.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med, 7*.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1-27.

Ouma, C. A. (2014). *Performance of cart-based value-added model against HLM, multiple regression, and student growth percentile value-added models.* PhD Thesis, Florida State University.

Potamites , L., Booker , K., Chaplin , D., & Isenberg , E. (2009). *Measuring school and teacher effectiveness in the EPIC Charter School Consortium-Year 2.* Washington, DC: USA: Mathematica Policy Research.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., Saxton, A., & Horn, B. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Schafer, W. D., Lissitz, R. W., Zhu, X., Hou, X., & Li, Y. (2012). Evaluating teachers and schools using student growth models. *Practical Assessment, Research & Evaluation, 7*(17).

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? *Economics of Education Review, 64*, 50-74.

Webster, W., & Mendro, R. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

Wei, H., Hembry, T., Murphy, D. L., & McBride, Y. (2012). *Value-added models in the evaluation of teacher effectiveness: A comparison of models and outcomes (Pearson Research Report).* New York, NY: Pearson.